DOCUMENT RESUME

ED 362 562 TM 020 617

AUTHOR	Dietel, Ron
TITLE	What Works in Performance Assessment? Proceedings of the CRESST Conference (Los Angeles, California, September 10-12, 1992). Evaluation Comment.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	93
CONTRACT	R117G10027
NOTE	25p.
PUB TYPE	Collected Works - Conference Proceedings (021) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cost Estimates; *Educational Assessment; Educational Research; Elementary Secondary Education; Evaluation Criteria; Generalization; *Performance; *Portfolios (Background Materials); *Psychometrics; Reliability; *Student Evaluation; Validity
IDENTIFIERS	*Alternative Assessment; Center for Research on Eval Standards Stu Test CA; *Performance Based Evaluation

ABSTRACT

The 1992 annual conference of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) was dedicated to explaining "What Works in Performance Assessment?" This report provides a synopsis of discussions by over 300 policymakers, researchers, and teachers. CRESST Co-Director Eva Baker summed up the present state of performance assessment in her opening remarks. Other researchers concurred with her warning that there is much that is not yet known about performance assessment. A beginning has been made, in that what is known about standardized tests and the misuse of their results has been defined. It is evident that state and federal interest in the use of performance assessments is growing. Research is being conducted into performance assessment fairness, particularly with regard to portfolios. The expected links among instruction, learning, and alternative assessment are being investigated. The CRESST validity criteria of transfer and generalizability have received substantial attention in studies of the technical aspects of performance assessment. The final validity criterion that CRESST requires of performance assessment focuses on cost and the resources needed to implement the new assessments in the classroom. What is known above all is that lots of performance assessments are being developed. Brief abstracts are given of 13 additional conference papers. An attachment lists and annotates the technical reports available from CRESST. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 362 562

UCLA's Center for the Study of Evaluation &
The National Center for Research on Evaluation, Standards, and Student Testing

EVALUATION COMMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Spring 1993

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J.C. BEER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

What Works in Performance Assessment?
Proceedings of the 1992 CRESST Conference

Ron Dietel, CRESST/UCLA[1]

"It is only education Luddites who would oppose new forms of assessment until everything is in place. These new assessments are too powerful a tool to stop using them. It's the business of researchers to be cautious—politicians are elected to act."

Policymaker statements from CRESST research

Performance assessment research appears to be caught in a law of supply and demand—plenty of demand from impatient customers who want to know if performance assessment actually "works," but too little "supply" in the form of answers from the research community about what works and what doesn't.

The 1992 annual CRESST assessment conference was dedicated to explaining "What Works in Performance Assessment." From September 10-12, 1992 over 300 policymakers, researchers, and teachers met on the UCLA campus to discuss what educators currently know about these new types of tests.

CRESST Co-director Eva Baker summed up the present state of performance assessment affairs in her opening conference remarks.

> "The policy and practitioner communities are acting," warned Baker, "with or without us. We no longer have the luxury of saying we don't have the answers yet but if you'll just hold on for four or five more years, our research will really be able to tell you what to do."

Baker cautioned against expecting too much from new assessment methods.

> "One of the things that worries many of us is the enormous hype that's been associated with performance assessment," said Baker. "It's better than superman, better than chocolate pecan pie, or the fastest, lightest computer notebook."

Other researchers concurred with Baker, including CRESST Co-director Robert Linn. Linn indicated that there is a lot more we *don't know* about performance

[1] *Special thanks to Joan Herman and Katharine Fry for their valuable suggestions to this article. Thanks also to the many presenters and discussants who shared their research at the 1992 CRESST conference.*

TM 020617

What Works in Performance Assessment?
The 1992 CRESST Conference

assessment than we *do know*, but that a conference focusing on what we have learned about performance assessment would contribute to the expanding base of performance assessment knowledge.

Linn suggested that a framework for the two-day meeting might include the CRESST validity criteria for performance assessment. Published in a 1991 CRESST technical report,[2] the CRESST criteria include:

- consequences of performance assessment;
- equity, including test fairness and opportunities for students to learn assessed knowledge skills;
- transfer and generalizability theory;
- content and curriculum quality including cognitive complexity, content quality, and content coverage;
- meaningfulness of performance assessments; and
- costs and efficiencies.

[2] *CSE/CRESST Technical Report 331*

A Beginning: What We Seem to Know About Standardized Tests

Test Scores Verses Performance

Although disagreeing on several points, many conference presenters agreed that there are serious problems with traditional standardized tests.

In her research with disadvantaged children, for example, Lily Wong Fillmore from the University of California, Berkeley, found a troubling discrepancy between results of standardized CTBS reading comprehension tests and actual student performance.

> "We were doing a performance assessment of language, of cultural adaptation to school, of how students were dealing with the problem of learning a language they did not know," said Wong Fillmore. "When we compared the [CTBS] test scores to what we were [performance] measuring, we found vast differences. With some of these kids, they were so free of English, that is, they would not speak a word of it—yet they did well in the CTBS reading comprehension test. Other students that we knew to be performing quite well," she added, "did very poorly on the CTBS."

This lack of correlation between standardized test scores and meaningful student performance has been noted by others. Standardized tests typically measure basic concepts and procedures, said Thomas Romberg, from the University of Wisconsin-Madison, but not in-depth understanding or student production of knowledge.

> "What they [standardized tests] measure, they measure well," noted Romberg, "but what they do not measure is of concern also. The main issue is to let people know that if these [standardized tests] are the instruments they are using, then basic concepts and procedures are the mathematics they are assessing, which is a small part of knowing and being able to use mathematics."

Improper Use of Standardized Tests Results

Other conference presenters had less negative views of standardized tests. H.D. Hoover, University of Iowa, suggested that the problem isn't with standardized tests themselves but the improper *use* of such assessments.

What Works in Performance Assessment?
The 1992 CRESST Conference

"There is no doubt," said Hoover, "the way [standardized] tests have been used has sometimes been horrible. I think this mostly has been brought about by mandated state assessment programs and using these tests for accountability in high-stakes situations for which they were never intended."

However, despite relatively widespread agreement about the problems of standardized tests and the utilization of their results, policymakers and the public will likely continue to use declining standardized test scores as a rallying cry for what's wrong with American education. And despite evidence that high stakes accountability "uses" corrupt the testing process, policymakers continue to have great faith in the power of assessment.

What We Know About Performance Assessment and Policy

State Interest in Assessment

State policymaker interest in assessment is growing, suggested Lorraine McDonnell from the University of California, Santa Barbara, because many policymakers view assessment as a lever of change. As part of a CRESST project, McDonnell is conducting an extensive investigation of the move towards new forms of assessment in Kentucky, California, Indiana, and North Carolina. She has found that state policymakers frequently support assessments for very different purposes.

In California, said McDonnell, the new performance-based California Learning Assessment System (CLAS) came about because of a rare consensus among the three state centers of education power: the governor, the legislature, and the state school superintendent. But each center had its own reasons for wanting new assessments.

"Governor Wilson would like to move to a system of merit-pay where teachers with high scoring students are rewarded," reported McDonnell. "One of the governor's aides told us: 'We could care less about authentic assessment—it costs more money and we don't know if it's any better. For us, having individual student scores is really powerful. It brings accountability into a system where it isn't there now. Parents can then say they don't want their child in Ms. Smith's classroom because they will have the [necessary] assessment information.'"

Offering a very different reason for the same new assessments was the California state legislature, primarily Senator Gary Hart, chair of the Senate Education Committee. According to McDonnell, Hart agreed to exchange "greater accountability in order to get greater autonomy for instruction and school operation. It was quid-pro-quo for having schools move to site-based management," said McDonnell.

The final policymaker in the game, Bill Honig, the former California Superintendent of Public Instruction, was "interested in assessments that are more congruent with the type of curriculum he espouses," added McDonnell, "assessments that measure real-world performance and will influence teaching."

The lesson, concluded McDonnell, is that test developers, schools, districts, researchers, and practitioners, will have to accommodate multiple and sometimes competing policymaker purposes that drive performance assessment development, implementation and use.

Federal Interest Grows

Meanwhile, the federal government has been looking at assessment as a lever for national educational reform. Several CRESST presenters suggested that momentum for national standards and national performance tests is rapidly growing.

> "Improving assessment became an issue [in recent years] not to improve assessment but because of the woeful state of education reported throughout the country," said Andrew Hartman, education policy coordinator for the Republican staff of education committee.

Agreeing with Hartman was Michael Feuer, Office of Technology Assessment.

> "The spirit in Washington has been tense," said Feuer. "National standards and national curriculum this last year have been prominent concepts in Washington and we seemed [at one point] to be moving towards national testing."

National standards are a likely reality, added Feuer. Yet fears about a single national test may lead to different assessments developed by individual states or clusters of states, with its own set of possible negative consequences. Policymakers, parents, the public, employers — everyone — will want to compare these different performance assessments to one another, despite warnings from the research community that precise comparisons may be technically impossible.

> "Comparing results from different types of high stakes assessments is questionable under most situations," said CRESST Co-director Bob Linn, "unless the assessments have been developed from very similar standards."

Furthermore, invalid comparisons may result in incorrect decisions about school or teacher performance, or worse yet, incorrect decisions about students.

What We Know About Performance Assessment and Fairness

More Issues Than Solutions

A second CRESST validity criterion, fairness, was addressed by several conference presenters who agreed that ensuring fairness for students who take performance assessments is at least as difficult as ensuring fairness for students who take standardized tests. They noted that the impact of new assessments on disadvantaged children could be severe if student opportunities to learn remain unequal.

> "If we put forth the worst-case scenario for African-American and Latino children," said CRESST researcher Linda Winfield, "where the instructional conditions are marginal, facilities are poor, where the actual assessment might be based on exercises or content that is totally foreign or the language is foreign, where the raters might be biased—then we have a situation where performance-based measures will be much worse than traditional measures in the form of standardized tests."

The language dependency of many alternative assessments is also troubling according to several conference participants, including CRESST Associate Director Joan Herman.

> "How do we separate language proficiency from content knowledge and thinking skills?" asked Herman. "The

problem is particularly acute for non-native speakers, who are disadvantaged by the many assessments requiring verbal fluency."

Lily Wong Fillmore also noted the reverse problem. She suggested that limited English speaking Asian students may understand less than what we assume and that such assumptions are cause for concern.

"Whether an assessment underestimates or overestimates a student's competence," said Wong Fillmore, "there are serious equity issues. If a test, whatever sort, favors individuals who are in fact not doing as well as you think, then what you get is a kind of *neglect* of the educational needs of those kids."

Wong Fillmore suggested that adequate attention and resources must be devoted to improving the skills of all children, regardless of cultural background.

Fairness in Portfolios

Research conducted in Pittsburgh by Paul LeMahieu, University of Delaware and the Delaware Department of Public Instruction, indicates that serious attention must be paid to whether or not portfolios are equitable for all students. During his presentation LeMahieu explained one of his portfolio studies:

"We examined two groups of students," said LeMahieu, "for whom we had access to both the full bodies of their work as well as the portfolios that resulted from their selections of work from that whole. One group scored higher on the portfolio selections. This group was made up of high achievers who were also predominantly white. The second group was a low achieving group, made up primarily of minority students. Their portfolios were rated lower than the full body of their work."

LeMahieu believes that the lower portfolio scores of the second group may have resulted because this group did not deeply understand the purposes of the portfolios, the standards against which their portfolio work would be measured, or that the students did not have the self-reflection skills necessary to assemble higher quality portfolios, ones that presented themselves more faithfully.

"Apparently the first group understood how their work would be judged," said LeMahieu, "and how to present themselves well with respect to the evaluative criteria in use. Moreover, they knew how to examine their work with a critical eye in compiling the portfolios that would represent them. The second group did not have access to these understandings. Obviously such knowledge and skills need to be the object of explicit instructions in order to avoid this potential source of bias."

Dennie Palmer Wolf from Project PACE, suggested that LeMahieu's research highlights the importance of instructional and assessment equity for portfolio use in the classroom. To compete on a level playing field, all students must have deep understandings of portfolio purposes, standards, and processes.

"There are all kinds of things about putting together your portfolio which we may be teaching to some students and not teaching to others," said Palmer Wolf. "We have a deep responsibility to think about these kinds of equity issues."

Ultimate Responsibility for Improved Education

Although equity-sensitive tests may significantly contribute to fairness in assessment, they cannot by themselves adequately solve the multitude of equity problems facing education today.

> "Improved assessment can contribute to improved education," said CRESST consultant Edmund W. Gordon, City University of New York, "but in the final analysis it is those of us who are responsible for teaching and learning that must ensure that adequate and equitable teaching occurs if effective learning is to be the result."

What We Know About The Promised Link Between Instruction, Learning, and Alternative Assessment

Cognitive theory provides a valuable framework for integrating performance assessments into the classroom. During the conference CRESST researcher Robert Glaser recommended:

> "Learning, instruction, and [performance] assessment should be one piece, a system of mutually interacting aspects of teaching. This system should be driven by the cognitive structures that are acquired by students as they achieve knowledge and skill in a subject matter."

Characteristics of knowledge typical of achieving students can be identified, explained Glaser, including how students structure, proceduralize and self-regulate knowledge for effective use. Glaser pointed out that some students jump into a problem or task without analyzing the nature of the problem, while higher achieving students form a model of the situation that enables them to generate possible approaches and select among various alternatives. Performance assessments should be capable of measuring such knowledge development processes, recommended Glaser.

Role of Teachers in Assessment

Glaser added that teachers play the pivotal role in this knowledge acquisition process. Other CRESST presenters echoed his feelings that teachers know their students better than just about anyone else.

> "Human teachers," said CRESST researcher Richard Snow from Stanford University, "are perhaps the most sensitive assessment device available for looking at student motivational and volitional behavior. Teachers can see it and sense it—it [student performance] is not always verbal."

Teachers must be actively involved in the entire assessment process if learning, instruction, and assessment are to become integrated, motivational factors in the classroom, said Jackie Cheong from the University of California, Davis. Cheong said that portfolios, such as the California Learning Record, enable teachers to understand key student learning processes. Integrating instruction and assessment, the California Learning Record is a portfolio assessment in which students' efforts are documented through structured observation by teachers.

Portfolios Integrate Learning, Instruction, and Assessment

A chief proponent of the value of portfolios in the learning, instruction and assessment process, Dennie Palmer Wolf has found that portfolios work best when they afford links across disciplines and are concerned *both* with high standards and with development. Teachers and schools should maintain portfolios on students for a period of years, said Palmer

Wolf, not just for a few months or a single school year. Performance tasks, whatever their nature, must be embedded into the curriculum.

> "Students must have time to think, to self-evaluate about their work," explained Palmer Wolf. "The [portfolio] technology is about doing cumulative work and having time in the school, in the curriculum, to reflect on that work."

Palmer Wolf's research from Arts Propel in Pittsburgh and Project PACE in four urban districts provides other "what we know" lessons that contribute to the role of portfolios in the learning, instruction and assessment process:

- Standards of performance should be made known to students as part of an overall school system that supports portfolio assessment.
- Portfolios must create a conversation between all the teachers in the school. Students should use portfolios to actually think, not merely record information.
- Portfolios should function as examples of student work that must be met prior to students' movement from one grade level to another. Portfolios should not live and die in the middle or elementary school; they must be maintained and passed on to the next higher institution of education. They should act as critical "passports" to the best educational options a student can locate and try.

What We Know About the Technical Aspects of Performance Assessment

The technical portion of the CRESST conference focused on what researchers have learned about task development, scoring, comparability, and moderation of performance assessment. As noted by several presenters, the CRESST validity criteria of transfer and generalizability have received substantial early attention and results are now becoming known.

What We Know About Task Development and Scoring

Many states or consortia of states or counties have embarked on an ambitious effort to develop performance assessment tasks. However, with few models to base their designs upon, developers have found this enterprise formidable and slow-going. Maryland is one state that has formed a successful consortium of counties which have pooled their resources to develop a performance system emphasizing thoughtful mastery of important tasks. But the development process has been problematic.

> "When one sees examples of finely crafted performance tasks, they look easy, but they are not," remarked Jay McTighe from the Maryland Assessment Consortium. "We found that task development is a long-term process and extraordinarily difficult work."

Other assessment developers have encountered similar hurdles. Lee Jones, who has been developing a new hands-on science performance assessment for the National Assessment of Educational Progress (NAEP), agreed with McTighe, noting that "the key thing we have learned is that this whole [development] process takes time, time, and more time."

Nevertheless, performance tasks are being developed and

some valuable lessons are arising from the process, including methods for reducing costs and resource needs. Eva Baker, along with other CRESST researchers, has developed a performance assessment model that appears applicable across a variety of subject areas and for a variety of topics. The result is a relatively cost-effective method for developing and scoring performance assessments.

CRESST Performance Assessment Model

Baker's model focuses on explanation skills. Originally applied in social studies/history, the model has been used to measure depth of understanding in science, mathematics, and geography. In the case of social studies/history, the assessment asks students to write extended essays on significant historical events, making use of prior knowledge and source materials such as the Lincoln-Douglas debates.

Some valuable "what we know" lessons have occurred from the research. The scoring system, for example, based on a comparison of expert and novice performances, showed that novices do not bring in external information (prior knowledge) to the assessment whereas experts do. Secondly, novices make some big mistakes—they misunderstand context and write in a very flat way.

> "In an attempt to be extremely comprehensive," said Baker, "they (novices) are very afraid to leave anything out. Experts on the other hand, write explanations that are very principle-oriented."

The analyses provided the scoring dimensions for the assessment: general impression of content quality, prior knowledge, principles or concepts, text detail, misconceptions, and argumentation. This general strategy of basing performance assessment scoring rubrics on differences between expert and novice performances shows promise for other assessments.

CRESST researchers also learned that by developing task specifications, blueprints for parallel tasks, they were able to reduce the number of tasks necessary to get relatively high reliability ratings. This important finding suggests that performance assessments may not require nearly the large number of tasks as suggested by other investigators.

What We Know About Group Assessment

Several states, such as Connecticut and California, are attempting to incorporate group assessment into their large-scale testing programs. One intention of such efforts is to use scores from group assessments as indicators of individual performance. However, a key technical question for such assessments is "To what extent do scores on a group assessment actually represent individual performance or knowledge?" A study by UCLA professor and CRESST researcher Noreen Webb sheds some light on this substantial technical question.

Webb gave two seventh-grade classes an initial mathematics test as a group assessment, where exchange of information and assistance was common. Several weeks later, she administered a nearly identical individual test to the same students where assistance was not permitted.

The results showed that some students' performance dropped significantly from the group assessment to the individual test. These students apparently depended on the resources of the group in order to get correct answers and when the same resources were not available during the individual test, many of the students were not able to solve the problems. Webb concluded:

> "Scores from a group assessment may not be valid indicators of some students' indi-

vidual competence. Furthermore, achievement scores from group assessment contexts provide little information about group functioning."

Webb's study suggests that states or school districts who intend to assign individual scores based on group assessments may want to seriously rethink their intentions.

Generalizability of Tasks and Assessment Methods

The number of tasks needed to ensure that assessments are reliable measures of student performance is one area where research has provided important results. Reported at the conference was valuable research on science performance tasks conducted by researchers at the University of California, Santa Barbara, including CRESST researchers Richard Shavelson and Gail Baxter.[3]

Looking for ways to reduce costs and task administration time, Shavelson, Baxter, and others designed a ***computer simulation*** of a hands-on performance task that was as close as possible to actual ***observations*** of student work. The researchers did everything they could to make the two methods — observations and computer simulations — comparable. But the results showed only a moderate correlation between the methods, even though they were painstakingly conceived, developed and administered.

> "This one [the two tasks] blew us away," said Shavelson. "There's actually a kid who got a [score of] one on the computer simulation task and a six when we observed his performance. And there's another kid who got a one when we observed his performance but got a six on the computer [task]. What this means," says Shavelson, "is that you get a different picture of kids' performance from two methods of measuring performance."

UCSB researchers also compared some short-answer responses and multiple-choice results, based on the same science tasks, to the computer simulation and observations of student performance. The only two tasks that appeared to be reasonably interchangeable were methods of direct observation and use of a notebook. The notebook required students to conduct the experiment and then report their procedures and results in a specific format.

> "The moral of the story is that most [performance assessment] methods are not interchangeable," concluded Shavelson.

This finding has key policy–making and cost implications: When attempting to make ***high-stakes*** decisions based on results from different assessments, many tasks and types of assessments may be needed in order to make valid generalizations of student performance.

What We Know About Comparing Performance Assessments

As previously mentioned, one proposal for a national assessment system would have clusters of states developing performance assessments matched to a national set of standards. CRESST Co-director Robert Linn, however, has strong questions about "if" and "how" the assessment community can make valid comparisons between different assessments developed in this manner. How will we know, for example, that these assessments are measuring the same thing? And based on the CRESST criteria, how will we know that the assessments are comparable in terms of their cognitive complexity, content quality, and content coverage?

[3] *Gail Baxter is now an assistant professor at the University of Michigan.*

During his CRESST presentation, Linn suggested that assessment comparability is one area with a plethora of unresolved issues including differences in tasks and administration conditions.

> "Administration of different assessments is just one part of this larger task comparability problem," said Linn. "How do we account for where, when and how long different performance tasks are administered and what instructional preparation children have had prior to taking the test?"

Linn said that task comparability must be considered in relation to two *types* of students—students who have never taken a performance test and students who have performance tests regularly embedded into their curriculum. "This issue has important equity and opportunity to learn implications," added Linn.

Technical Lessons from the United Kingdom—Moderation

The United Kingdom is considered well ahead of the United States in the development, use, and scoring of performance assessments. Many of the comparability issues mentioned by Bob Linn are ones that the U.K. educational system has had to address. CRESST presenter Desmond Nuttall from the University of London explained the comparability problems the British have encountered where performance assessments have been tied to national standards.

> "A major issue for us," said Nuttall, "is whether a grade A in Southampton has the same meaning, the same utility, the same standard, as it does in Newcastle. We also have to face the issues of comparability over time and comparability over different subjects," he added.

Nuttall noted at least one other technical problem that may have implications in the United States. Grade inflation.

> "In 1988 we introduced a new examination based on performance assessment," said Nuttall, "and in the years since, we've seen grades improve dramatically. In 1988, some 42% of the students achieved a grade C or better in the examination. This year the figure has risen from 42% to 51%. Everyone was congratulating themselves on the great success of education in raising the performance of students until our secretary of state revealed a report which suggests that there were a lot of fallibilities in human judgment that had gone into the assessment [scoring] and that it was a phenomenon well-known to you [Americans], grade inflation, rather than a real improvement of standards. Not everyone agrees with the Education Secretary, John Patten, but he proceeded to tighten the system."

In response to such problems, Nuttall said the United Kingdom turned to moderation to verify performance assessment scoring methods. According to Nuttall:

> "Moderation is the basic quality assurance mechanism we use to make sure that assessments not only meet the national content and performance standards, but also meet requirements of validity, reliability, and equity."

The United Kingdom now uses teachers and curriculum consultants as moderators for their national assessments. Regularly meeting and reviewing student work, teachers reach consensus on questions of task comparabil-

ity, standards, and scoring procedures. In cases where the work cannot be brought to a central site, the teachers visit individual schools to confer and evaluate student work and attempt to moderate student performances against the national standards. In addition to increasing the reliability of the assessments, the process itself has had a very positive effect on teachers.

> "The model of every teacher as a moderator," said Nuttall, "is a powerful device for professional development of teachers—giving them access to different ways of both assessing students and of setting suitable activities for students, preparing them for assessment."

Nuttall noted that this method may not meet the traditional technical criteria of validity and reliability, but that the inclusion of teachers in the moderation process has brought important gains in supporting the comparability of assessment and scoring and in enriching teachers' professional skills.

At least one other CRESST presenter reported similar benefits of getting teachers deeply involved in the entire assessment formula. Jay McTighe, Maryland Assessment Consortium, said:

> "Getting teachers involved early on in the guts of [Maryland assessment] development was very relevant to the process. We have learned that working with others to develop performance assessment and scoring instruments is one of the most powerful forms of professional development possible."

Ultimately what researchers and others learn from the development of performance assessments within the United States and from other countries will address the technical validity criteria of performance assessment tasks.

What We Know About Performance Assessment Costs and Resources

The final CRESST validity criterion focuses on the costs of performance assessment and the resources needed to implement new assessments in the classroom. Although there are few details about specific costs and resource requirements associated with performance assessments, anecdotal evidence indicates that such assessments are expensive, time-consuming, and resource-intensive.

Resource Needs

CRESST presenters uniformly agreed that developing and implementing classroom performance assessments places a tremendous burden on teachers. Teachers need extra time and extensive professional development, both of which are usually lacking in most schools today, if they are to become involved and committed to the assessment reform process. For example, CRESST researcher Charlotte Higuchi from Farmdale Elementary School suggested during her presentation:

> "We need time at the classroom level! Two weeks before school, to think, to plan, to write, to learn, to innovate, to design [performance assessments]."

Higuchi added that performance assessments require extra time to conference with children and parents, to review anecdotal notes on students, and especially "time for teachers to think."

Professional development is another agreed-upon resource prerequisite. CRESST researchers Maryl Gearhart and Shelby Wolf, for example, found that

teachers involved in a portfolio program required substantial, continuous professional development to help them implement standards for student writing. Finding that teachers did not critically assess their students' portfolio work and did not fully understand what constitutes quality writing standards, Wolf conducted a series of workshops to help teachers discern elements of good writing and then built an assessment rubric founded on these elements. Gearhart stressed that teachers must have substantial knowledge of a subject before they can be expected to be good assessors of it. She noted, however, that few schools or districts are able to fund this type of comprehensive development program.

Opportunity Costs

There are other expenses associated with performance assessments, including opportunity costs. Within a fixed school day, any added program means the loss of something else. Add portfolios and you might have to lose computer training. Vermont, for example, implemented portfolios as a statewide assessment during the 1991-92 school year, focusing on mathematics and writing skills. According to a CRESST evaluation, in order to implement portfolio programs in these two topics, Vermont teachers cut back their teaching of other subjects.

> "Performance assessment costs," said CRESST researcher and RAND social scientist Daniel Koretz who evaluated the Vermont program. "And the financial costs I think are not the largest," he added, "there is a cost in [loss of] content coverage. What will happen when they [Vermont teachers] have four [portfolio] subjects two years down the road, I don't know."

Solutions for High Costs and Resource Needs

CRESST researcher Lorrie Shepard from the University of Colorado, Boulder, suggested an alternative for classroom teachers who don't have the time to develop performance assessments on their own but who are dissatisfied with assessments imposed by others.

> "Steal half of the things mentioned at this conference and look at them in detail," urged Shepard to teachers attending the conference. Many examples not used for secure assessments are in the public domain. Look at the tasks and start collecting different samples of performance assessment. See what they look like and what the scoring criteria are that make sense."

Shepard recommended that after teachers collect enough tasks, they should adapt these performance assessments to their own schools and classrooms and then learn to develop their own.

Presenter Thomas Payzant, nominee for Assistant Secretary for Elementary and Secondary Education and former Superintendent of San Diego City Schools, was philosophical about the various time and cost burdens discussed by others.

> "Before we become too hard on assessment," said Payzant, "remember that if we do curriculum development, it's hard and takes a lot of energy. If we really focus on teaching strategies and how kids learn, it's hard, time consuming and costly, and takes a lot of energy. So why should we expect the assessment effort to be any different? Perhaps we can get some economies of energy and scale by doing the three simultaneously," recommended Payzant.

What Works in Performance Assessment?
The 1992 CRESST Conference

Dan Resnick who is directing assessment development for the New Standards Project has found that the most efficient way to develop performance assessments is through a partnership of organizations who share costs. Seventeen states and six urban districts have joined the New Standards Project, the largest single group of states and districts across the country developing performance assessments matched to specific standards.

Meanwhile, as already mentioned, Maryland has successfully pooled its statewide resources and created an assessment consortium involving all twenty-four school districts in the state. Creative assessments will require creative solutions to solve significant time and resource constraints.

Worthwhile Results With Resources In Place

The good news is that when adequate resources are in place, assessment reform appears to happen. Presenter Elizabeth Rogers from Charlottesville City (VA) Public Schools noted that once her school district met teachers' needs—in terms of professional development, dissemination of research, and instructional support—change occurred.

> "Teachers began to invent, to talk to other teachers, to read research," said Rogers, "and they created sophisticated [performance] methods for assessing students."

The admonition is that anyone who thinks that assessment reform can occur without additional resources is likely to find that few classroom changes are actually implemented. But when resource needs are foreseen and met, and teachers are part of the entire process, change is not only possible, but significant.

What Else We Know About Performance Assessment

What we know is that lots of performance assessments are being developed. CRESST presenters shared their efforts in the areas of mathematics, science, literacy, social studies, workforce readiness, and several other topics, such as portfolios and multidisciplinary assessment. The following comments from various presenters contributed to the growing knowledge of "what works" in performance assessment.

Effects of Performance Assessment

- One positive effect of performance assessment is that learning and assessment is now a continuous process: Students have to re-do [performance] tasks.

  Melody Ulen,
  Littleton High School, Denver
  Performance Assessments in Science

- On balance, people [Vermont teachers and principals] were positive about the impact on instruction. Some people were practically euphoric. Teachers who everyone thought were the least likely to change their instruction were, in fact, finally changing.

  Daniel Koretz,
  CRESST/The RAND Corporation
  Models for Collaborative Assessment Development

- Performance assessment does not take away from instruction but instead offers students another opportunity to learn.

  Gail Baxter,
  University of Michigan
  Performance Assessments in Science

- Teachers describe significant shifts in their instructional and assessment practices. To a large extent they cannot and do not separate the two.

Elizabeth Rogers,
Charlottesville City Public Schools
Performance Assessments in Literacy

- In the national evaluation of performance assessments, teachers reported the [scoring] agreement that they were able to reach as one of the most important and useful parts of the whole performance assessment paraphernalia. The process also helped to insure more uniform assessment among teachers.

Desmond Nuttall,
University of London
Lessons From Performance Assessment in the United Kingdom

- Authentic assessment in social studies seems to lead students to become involved in a topic, providing students a deeper understanding of social issues, and a greater comprehension of the interconnections of specific historical periods.

Cris Gutierrez,
Jefferson High School, Los Angeles
Performance Assessments in Social Studies

- Assessment does drive school instruction. As one teacher involved in piloting a science task remarked, "If this is the way you are testing, then this is the way we are going to teach. This will really impact what we do in our classroom."

Kathy Comfort,
California Department of Education
Performance Assessments in Science

Many Remaining Challenges

- Unfortunately, the performances on the open-ended items [performance assessments in science] have very thin results indeed. This recurrent problem may be attributable to students' inexperience with performance assessments or [the possibility] that many students just aren't very good writers.

Darrell Bock,
CRESST/University of Chicago
Science Performance Assessments

- We are a lot less further along about delivery standards than we are about content and performance standards. Delivery standards are criteria to judge whether states, school systems, schools, and classrooms are providing an education that will enable students to achieve those [content and performance] standards. In essence, the burden should be on the system, not on the students, for educating children. Otherwise, performance standards are not fair.

Hilda Borko,
CRESST/University of Colorado
Service Delivery Standards

- In scoring our field tests in reading, mathematics, and social studies, we found that students had difficulty writing about specific content areas. They could often arrive at an answer but be unable to explain how they arrived at the answer or why it was the correct one. They are not accustomed to justifying or explaining—they are accustomed to recall and formulas.

Daisy Vickers,
North Carolina State Department of Instruction
Multidisciplinary Assessments

- Findings from the QUASAR project indicate that

What Works in Performance Assessment? The 1992 CRESST Conference

student performance across different mathematics tasks is inconsistent. Consequently, if student-level scores are of interest, more than nine tasks may be required to obtain reliable results of student performance.

Suzanne Lane,
University of Pittsburgh
Performance Assessment in Mathematics

Findings About Prompts and Scoring

- One of the tricks in multidisciplinary assessment is that you can give students a single task, but use different kinds of rubrics for scoring, depending on what you are interested in measuring. For example, we asked students in the Humanitas multidisciplinary program to write essays on a topic integrating their knowledge of history, literature, and the arts, and then we scored the essays according to both writing quality and subject matter understanding.

Pamela Aschbacher
CRESST/UCLA
Multidisciplinary Assessments

- Many tests used for accountability, such as NAEP, have high stakes for administrators and no stakes for students. A key question is how to motivate students when the test does not count. Our research indicates that financial incentives increase test performance but non-financial incentives do not. These findings were particularly true for eighth graders on easier NAEP items.

Harold O'Neil, Jr.
CRESST/
University of Southern California
NAEP Motivation Study

In Conclusion

The 1992 annual CRESST conference synthesized much of the current knowledge of what researchers, teachers, and assessment policymakers believe "works" in performance assessment. Although much has been discovered in the last few years, the current performance assessment movement is still in its infancy, and the demand for knowledge about what "works" in performance assessment will likely exceed the supply for some time to come.

Portfolio Assessment Videotape!

Join the CRESST research staff, including Eva Baker and Maryl Gearhart, in "Portfolio Assessment and High Technology." This 10-minute production, made in 1992, examines key issues of portfolio assessment including:

- Student use of portfolios in the classroom;
- Selecting students' best pieces of classroom work;
- Involvement of parents in the portfolio process;
- Use of technology to promote good writing;
- Electronic student portfolios.

This videotape will be useful to school districts, principals and teachers interested in building their own portfolio programs, as well as researchers who want more information about the latest CRESST research programs.

The cost of "Portfolio Assessment and High Technology" is $10.00 and may be ordered on page 24.

CRESST/CSE Technical Reports

The following technical reports are now available by calling (310) 206-1532. Or you may fill in the order form on page 24 and mail to UCLA/CRESST, Graduate School of Education, 405 Hilgard Avenue, Los Angeles, CA 90024-1522

NEW!

Performance-Based Assessment and What Teachers Need

Charlotte Higuchi

CSE Technical Report 362, 1993 ($4.00)

Arguing that effective implementation of performance assessments in the classroom requires systemic reform of the teaching profession and of the school systems in which they work, teacher Charlotte Higuchi discusses the criteria that will result in improved classroom assessment. She suggests that alternative assessments offer teachers a critical tool for understanding their children.

"Multiple-choice tests eliminate teacher judgment in the assessment process," says Higuchi, "and are frequently not aligned with the instructional program. In contrast, performance-based assessments *are* individual or collective teacher judgments. They give rich, detailed information as to what students can and cannot do, and therefore enable teachers to plan instruction based on student needs.

To help teachers develop and implement their own performance assessments and to become teacher-researchers, Higuchi urges school districts to provide the following minimum resources:

- Time to think, to learn, to write, to collaborate, to analyze, to plan, and to create new forms of assessment;
- Work space including desks, chairs, file cabinets, and storage cabinets for equipment;
- Computers and printers, a phone, and a fax;
- Duplicating services to copy student work for portfolios and assessment records;
- Clerical support to type correspondence, order materials, and maintain records;
- An onsite library with journals from professional organizations, the latest books on education, and a media center.

"Full implementation of performance-based assessments," says Higuchi, "demands that teachers constantly discuss student performance, standards of performance, and how to change the instructional program to improve that performance."

STILL NEW!

Sampling Variability of Performance Assessments

Richard Shavelson, Xiaohong Gao and Gail Baxter

CSE Technical Report 361, 1993 ($4.00)

The authors of this study examined the cause of measurement error in a number of science performance assessments. In one part of the study, 186 fifth- and sixth-grade students completed each of three science tasks: an experiment to measure the absorbency of paper towels; a task that measured students' ability to discover the electrical contents of a black mystery box; and a task requiring students to determine sow bugs' preferences for various environments (damp vs. dry, light vs. dark).

The researchers found that the measurement error was largely due to task sampling variability. In essence, student performance var-

CRESST/CSE Technical Reports

ied significantly from one task sample to another.

Based on their study of both science and mathematics performance assessments, the authors concluded that "regardless of the subject matter (mathematics or science), domain (education or job performance) or the level of analysis (individual or school), large numbers of tasks are needed to get a generalizable [dependable] measure of performance."

In another part of the study, the researchers evaluated the *methods* in which students were assessed on several of the same experiments including:

- a notebook—a method in which students conducted the experiment, then described in a notebook the procedures they followed and their conclusions;
- computer simulations of the tasks; and
- short-answer problems where students answered questions dealing with planning, analyzing or interpreting the tasks.

The notebook and direct observations were the only methods that appeared to be fairly interchangeable. The results from both the short-answer problems and the computer simulations were disappointing.

Increasing the number of tasks is costly and time consuming, conclude the authors. But they warn that trying to explain away technical problems is dangerous.

CRESST Performance Assessment Models: Assessing Content Area Explanations
Eva Baker, Pamela Aschbacher, David Niemi, and Edynn Sato, 1992 ($10.00)

This assessment model, based on a highly contextualized history performance task, requires students to engage in a sequence of assessed steps, including an initial evaluation of their relevant background knowledge of the particular historical period. Students write an extended essay that explains the positions of the authors of the original text materials, such as the Lincoln-Douglas debates, and draw upon their own background knowledge for explanation. The essay scoring rubric consists of six dimensions: a General Impression of Content Quality scale, and five analytic subscales.

Included in the handbook are: background information on the CRESST performance-based assessment, examples of assessments for secondary-level history and chemistry, and specifications for duplicating the technique with other topics and subject matter areas. The rater training process, scoring techniques, and methods for reporting results are described in detail.

Raising the Stakes of Test Administration: The Impact on Student Performance on NAEP
Vonda L. Kiplinger and Robert L. Linn
CSE Technical Report 360, 1993 ($4.00)

The National Assessment of Educational Progress (NAEP) test has been accused of underestimating student achievement because this "low-stakes" assessment has no consequences for students, their teachers, or their schools. In contrast, "high-stakes" tests—those assessments that have serious consequences for students, teachers, and schools—are assumed to motivate greater student performance because of the positive or negative consequences (such as college entrance) associated with student performance on the test.

The purpose of this study was to investigate whether differences in test administration conditions

and presumed levels of motivation created by the different testing environments affect student performance on the NAEP test. The testing conditions studied were the "low-stakes" environment of the current NAEP administration and a "higher-stakes" environment typified by many state assessment programs.

The results of the study lead to the conclusion that estimates of achievement from NAEP would not be substantially higher if the stakes were increased to the level associated with a "higher-stakes" test.

Issues in Innovative Assessment for Classroom Practice: Barriers and Facilitators

Pamela Aschbacher

CSE Technical Report 359, 1993 (**$4.50**)

As proven by the British experience, we cannot assume that new innovative assessments will be immediately understood and embraced by American teachers. Implementing performance assessments may demand new roles for teachers and students and require a radical paradigm shift among educators—from a focus on content coverage to outcomes achieved.

This paper, utilizing an action research approach, describes the findings of CRESST researchers who observed, interviewed, and surveyed teachers involved in implementing alternative assessments into their classrooms. Probably the most fundamental barrier to developing and implementing sound performance assessments was the pervasive tendency of teachers to think about classroom activities rather than student outcomes. Teachers who used portfolios, for example, focused on what interesting activities might be documented in the portfolios rather than what goals would be achieved as a result of these instructional activities.

The study revealed other basic barriers in the development and implementation of alternative assessments, including teacher assessment anxiety, lack of teacher time and training, and teachers' reluctance to change.

Writing What You Read: Assessment as a Learning Event

Shelby Wolf and Maryl Gearhart

CSE Technical Report 358, 1993 (**$4.00**)

This report focuses on the central role of teachers' interpretative assessments in guiding the growth of young writers. The teacher serves as critical reader and responder, providing commendations and recommendations for further growth. But the teacher is not the only expert. Students too are encouraged to participate in assessment dialogues, reflecting, analyzing, and contributing to their growth.

The authors of this report propose a new scheme to guide teachers' and students' reflection. Focusing on narrative criticism and composition, the scheme is based on eight components of narrative: genre, theme, characters, setting, plot, point of view, style, and tone.

Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress

Daniel Koretz, Elizabeth Lewis, Tom Skewes-Cox, and Leigh Burstein

CSE Technical Report 357, 1992 (**$4.00**)

Non-response to test items on the National Assessment of Educational Progress has been a concern for some time, particularly in the case of mathematics. Until recently, the primary concern has been "not-reached" items—that is, items not answered because the student failed to complete the test—as opposed to omitted or skipped items.

The study examined patterns of non-response in the three age/grade groups (age 9/grade 4, age

CRESST/CSE Technical Reports

13/grade 8, and age 17/grade 12) included in the 1990 assessment of mathematics.

The results showed that overall omit rates were modest in grades 4 and 8, and not-reached rates were greatly reduced from 1986 levels. Differences in non-response between white and minority students were less severe than they first appeared when adjusted for apparent proficiency differences. Gender differences in omit rates were infrequent.

Nonetheless, the results provide grounds for concern. Omit rates were high for a subset of open-ended items, and the proportion of items with high omit rates in grade 12 was substantial. The omit-rate differentials between white and minority students, especially for open-ended items, are troubling and will likely become more so as the NAEP continues to increase its reliance on such items. Taken together, these results suggest the need for routine but focused monitoring and reporting of non-response patterns.

Latent Variable Modeling of Growth With Missing Data and Multilevel Data

Bengt Muthén

CSE Technical Report 356, 1992 ($2.50)

This paper describes three important methods of multivariate analysis which are not always thought of in terms of latent variable constructs, but for which latent variable modeling can be used to great advantage. These methods are: random coefficients describing individual differences in growth; unobserved variables corresponding to missing data; and variance components describing data from cluster sampling. The methods are illustrated using mathematics achievement data from the National Longitudinal Study of America Youth.

The Reliability of Scores From the 1992 Vermont Portfolio Assessment Program

Daniel Koretz, Brian Stecher, and Edward Deibert

CSE Technical Report 355, 1993 ($3.00)

A follow-up report to the same study (CSE Report 350), this report presents CRESST's findings about the reliability of scores from the Vermont portfolio assessment program. In this component, the researchers focused not on the program's impact as an educational intervention, but rather on its quality as an assessment tool.

The "rater reliability"—that is, the extent of agreement between raters about the quality of students' portfolio work--was on average low in both mathematics and writing. However, reliability varied, depending on subject, grade level, and the particular scoring criterion, and in a few instances it could be characterized as moderate. The overall pattern was one of low reliability, however, and in no instance was the scoring highly reliable.

Although it may be unrealistic to expect the reliability of portfolio scores to reach the levels obtained in standardized performance assessments, the Vermont portfolio assessment reliability coefficients are low enough to limit seriously the uses of the 1992 assessment results. The report concludes with an analysis of issues that need to be considered in improving the technical quality of the assessment.

CRESST/CSE Technical Reports

Assessment of Conative Constructs for Educational Research and Evaluation: A Catalogue

Richard Snow and Douglas Jackson

CSE Technical Report 354, 1992 ($8.00)

In recent years, an overabundance of psychological constructs and their associated measures have been presented by educational researchers and program evaluators. Among the most interesting and potentially useful of these constructs are those reflecting motivational and volitional aspects of human behavior, called "conative constructs." Among the constructs in this category are: need for achievement and fear of failure, beliefs about one's own abilities and their development, feelings of self-esteem and self-efficacy, attitudes about particular subject-matter learning, and many others.

This catalogue brings together in one place those conative constructs that seem most promising as useful for future research and evaluation work in education. For each catalogued construct, the authors provide a brief review covering construct definition, theoretical base, assessment procedures, references, and where possible, study abstracts evaluating assessment instruments or otherwise bearing on appropriate construct validation.

The Apple Classrooms of Tomorrow^{sm}: The UCLA Evaluation Studies

Eva L. Baker, Maryl Gearhart, and Joan L. Herman

CSE Technical Report 353, 1993 ($3.50)

The Apple Classrooms of Tomorrow^{sm} (ACOT) project was initiated in classrooms at five school sites in 1985 as a program of research on the impact of interactive technologies on teaching and learning. While the project has expanded over time to encompass a larger and more diverse set of efforts, key components at all sites were the provision of high technology access, site freedom to develop technology-supported curriculum and pedagogy as appropriate to site goals, and the resulting study of what happens when technology support is readily available to students and teachers.

Four basic questions guided the evaluation:

1. What is the impact of ACOT on students?
2. What is the impact of ACOT on teachers' practices and classroom processes?
3. What is the impact of ACOT on teachers professionally and personally?
4. What is the impact of ACOT on parents and home life?

This report summarizes findings from 1987 through 1990.

Collaborative Group Versus Individual Assessment in Mathematics: Group Processes and Outcomes

Noreen Webb

CSE Technical Report 352, 1993 ($4.00)

This study asked the question: "To what extent do scores on a group assessment actually represent individual performance or knowledge?" Researcher Noreen Webb gave two seventh-grade classes an initial mathematics test as a group assessment, where exchange of information and assistance was common. Several weeks later, she administered a nearly identical individual test to the same students where assistance was not permitted.

The results showed that some students' performance dropped significantly from the group assessment to the individual test. These students apparently depended on the resources of the group in order to get correct an-

CRESST/CSE Technical Reports

swers and when the same resources were not available during the individual test, many of the students were not able to solve the problems. "Scores from a group assessment," said Webb, "may not be valid indicators of some students' individual competence. Furthermore, achievement scores from group assessment contexts provide little information about group functioning."

Educational Assessment: Expanded Expectations and Challenges

Robert Linn

CSE Technical Report 351, 1992 ($3.50)

"Educational policymakers are keenly interested in educational assessment," says Robert L. Linn in his 1992 Thorndike Award address to the American Psychological Association. Linn points to the various attractions that assessments have for policy makers who frequently think of assessment as a "kind of impartial barometer of educational quality." But assessments are frequently used for two questionable purposes, notes Linn, first, to point out the declining quality of American education and, secondly, as an instrument of educational reform. "Such greatly expanded, and sometimes unrealistic, policymaker expectations" he says, "together with the current press for radical changes in the nature of assessments, represent major challenges for educational measurement." Linn concludes his remarks by saying that the measurement research community must make sure that the consequences for any new high-stakes performance assessment system are better investigated than they were for previous assessment reforms.

The Vermont Portfolio Assessment Program: Interim Report on Implementation and Impact, 1991-92 School Year

Daniel Koretz, Brian Stecher, and Edward Deibert

CSE Technical Report 350, 1992 ($6.00)

Vermont is the first state to make portfolios the backbone of a statewide assessment system. Daniel Koretz, Brian Stecher, and Edward Deibert, the authors of this CRESST/RAND report, have been evaluating the Vermont portfolio program for almost two years. The researchers found that support for the Vermont portfolio program, despite tremendous demands on teacher time, is widespread. "Perhaps the most telling sign of support for the Vermont portfolio program," write the authors, "is that [even in the pilot year] the portfolio program had already been extended beyond the grades targeted by the state."

An interesting instructional phenomenon was that over 80% of the surveyed teachers in the Vermont study indicated that they had changed their opinion of students' mathematical abilities based upon their students' portfolio work. In many cases, teachers noted that students did not perform as well on the portfolio tasks as on previous classroom work. This finding, supported by other performance assessment research, suggests that portfolios may give teachers another assessment tool that appears to broaden their understanding of student achievement.

Design Characteristics of Science Performance Assessments

Robert Glaser, Kalyani Raghavan, and Gail Baxter

CSE Technical Report 349, 1992 ($3.00)

Part of a long-range goal to investigate the validity of reasoning and problem-solving assessment tasks in science, this report describes progress in analyzing several science performance assessment projects. The authors discuss developments from Connecticut's Common Core of Learning Assessment Project, the California Assessment Program,

and the University of California, Santa Barbara/California Institute of Technology research project "Alternative Technologies for Assessing Science Understanding." The analysis framework articulates general aspects of problem-solving performance, including structured, integrated knowledge; effective problem representation; proceduralized knowledge; automaticity; and self-regulatory skills.

Accountability and Alternative Assessment
Joan Herman
CSE Technical Report 348, 1992 ($4.00)

Despite growing dissatisfaction with traditional multiple-choice tests, national and state educational policies reflect continuing belief in the power of good assessment to encourage school improvement. The underlying logic is strong. Good assessment sets meaningful standards, and these standards provide direction for instructional efforts and models of good practice. But are these reasonable assumptions? How close are we to having the good assessments that are required?

This report summarizes the research evidence supporting current beliefs in testing, identifies critical qualities that good assessment should exemplify, and reviews the current state of the research knowledge on how to produce such measures.

Benchmarking Text Understanding Systems to Human Performance: An Exploration
Frances Butler, Eva Baker, Tine Falk, Howard Herl, Youngchee Jang, and Patricia Mutch
CSE Technical Report 347, 1991 ($5.00)

Benchmarking in the context of this report means comparing the performance of intelligent computer systems to the performance of humans on the same task. The results of this report support the belief that we can compare *system* performance to *human* performance in a meaningful way using performance-based measures. This study provides direction for researchers who are interested in a methodology for assessing intelligent computer systems.

More Reports

For a list of over 150 CRESST/CSE technical reports, monographs and products, please write to: CRESST/UCLA, Graduate School of Education, 405 Hilgard Avenue, Los Angeles, CA 90024-1522. Or call Kim Hurst at (310) 206-1532.

CRESST Conference

Registration forms for the 1993 CRESST conference will be on their way soon. Please save the dates September 13-14, 1993 on your calendar for this special event!

UCLA's Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing
Eva L. Baker, Co-director
Robert L. Linn, Co-director
Joan L. Herman, Associate Director
Ronald Dietel, Editor
Katharine Fry, Editorial Assistant

The work reported in this publication was supported under the Educational Research and Development Center Program cooperative agreement number R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Monographs and Resource Papers

MONOGRAPHS

Assessing Student Achievement: A Profile of Classroom Practices
Dorr-Bremme & Herman
CSE Monograph 11, 1986 ($11.00)

Evaluation in School Districts: Organizational Perspectives
Bank & Williams (Editors)
CSE Monograph 10, 1981 ($7.50)

Values, Inquiry and Education
Gideonse, Koff & Schwab (Editors)
CSE Monograph 9, 1980 ($11.00)

Toward a Methodology of Naturalistic Inquiry in Educational Evaluation
Guba
CSE Monograph 8, 1978 ($4.50)

The Logic of Evaluative Argument
House
CSE Monograph 7, 1977 ($4.50)

Achievement Test Items—Methods of Study
Harris, Pearlman & Wilcox
CSE Monograph 6, 1977 ($4.50)

RESOURCE PAPERS

Improving Large-Scale Assessment
Aschbacher, Baker & Herman
CSE Resource Paper 9 ($10.00)

Improving Opportunities for Underachieving Minority Students: A Planning Guide for Community Action
Bain & Herman
CSE Resource Paper 8 ($11.00)

Designing and Evaluating Language Programs for African-American Dialect Speakers: Some Guidelines for Educators
Brooks
CSE Resource Paper 7 ($2.00)

A Practical Approach to Local Test Development
Burry, Herman & Baker
CSE Resource Paper 6 ($3.50)

Analytic Scales for Assessing Students' Expository and Narrative Writing Skills
Quellmalz & Burry
CSE Resource Paper 5 ($3.00)

Criteria for Reviewing District Competency Tests
Herman
CSE Resource Paper 4 ($2.00)

Issues in Achievement Testing
Baker
CSE Resource Paper 3 ($2.50)

Evaluation and Documentation: Making Them Work Together
Burry
CSE Resource Paper 2 ($2.50)

An Introduction to Assessment and Design in Bilingual Education
Burry
CSE Resource Paper 1 ($3.00)

FOLD AND SECURE

Place Postage Here

CSE/CRESST
UCLA Graduate School of Education
405 Hilgard Avenue
Los Angeles, CA 90024-1522

Order Form

Attach additional sheet if more room is needed. Form is pre-addressed on reverse.

CSE Reports/Monographs/Resource Papers/Videotapes

Report Number	Title	Number of copies	Price per copy	Total Price

POSTAGE & HANDLING
(Special 4th Class Book Rate)

Subtotal of	$0 to $10	add $1.50
	$10 to $20	add $2.50
	$20 to $50	add $3.50
	over $50	add 10% of Subtotal

ORDER SUBTOTAL _____

POSTAGE & HANDLING (scale at left) _____

California residents add 8.25% _____

TOTAL _____

Orders of less than $10.00 must be prepaid

Your name & mailing address—please print or type:

_____

_____

_____

❑ Payment enclosed ❑ Please bill me

❑ I would like to receive free copies of the *CRESST Line* and *Evaluation Comment* publications.

UCLA
CSE/CRESST
Graduate School of Education
405 Hilgard Ave.
Los Angeles, California 90024-1522

NONPROFIT ORG.
U.S. POSTAGE
PAID
U.C.L.A.

ADDRESS CORRECTION REQUESTED (EE 72)

Larry Rudner
FRIC
3333 K Street, NW, Suite 200
Washington DC 20007